Multivariate and multiscale dependence in the global climate system revealed through complex networks

Karsten Steinhaeuser · Auroop R. Ganguly · Nitesh V. Chawla

Received: 19 April 2011 / Accepted: 23 June 2011 / Published online: 30 June 2011

Abstract A systematic characterization of multivariate dependence at multiple spatio-temporal scales is critical to understanding climate system dynamics and improving predictive ability from models and data. However, dependence structures in climate are complex due to nonlinear dynamical generating processes, long-range spatial and long-memory temporal relationships, as well as low-frequency variability. Here we utilize complex networks to explore dependence in climate data. Specifically, networks constructed from reanalysis-based atmospheric variables over oceans and partitioned with community detection methods demonstrate the potential to capture regional and global dependence structures within and among climate variables. Proximity-based dependence as well as long-range spatial relationships are examined along with their evolution over time, yielding new insights on ocean meteorology. The tools are implicitly validated by confirming conceptual understanding about aggregate correlations and teleconnections. Our results also suggest a close similarity of observed dependence patterns in relative humidity and horizontal wind speed over oceans. In addition, updraft velocity, which relates to convective activity over the oceans, exhibits short spatiotemporal decorrelation scales but long-range dependence over time. The multivariate and multi-scale dependence patterns broadly persist over multiple time windows. Our findings motivate further investigations of dependence structures among observations, reanalysis and model-simulated data to enhance process understanding, assess model reliability and improve regional climate predictions.

Keywords Complex networks · Correlation · Teleconnections · Reanalysis data · Ocean meteorology

1 Introduction

Developing a better understanding of the climate system and producing enhanced predictive insights are often confounded by complex dependence structures among climate variables including long-range spatial dependence, long-memory temporal processes, interactions at multiple scales, and nonlinearity of the underlying processes and relationships (Goddard et al. 2001; Hoerling et al. 2010). Complex networks have been motivated in climate to understand attributes of large-scale dynamics, for example, correlations within variables as a function of geographical proximity and teleconnections (Gozolchiani et al. 2008; Steinhaeuser et al. 2010b; Tsonis et al. 2006), inherent predictability of climate over oceans (Steinhaeuser et al. 2010a, 2010; Tsonis et al. 2006) and relations among ocean-based oscillators (Gozolchiani et al. 2008; Tsonis et al. 2006; Tsonis and Swanson 2008; Yamasaki et al.

Electronic supplementary material The online version of this article (doi:10.1007/s00382-011-1135-9) contains supplementary material, which is available to authorized users.

K. Steinhaeuser · A. R. Ganguly (✉)
Geographic Information Science and Technology Group,
Computational Sciences and Engineering Division,
Oak Ridge National Laboratory, 1 Bethel Valley Rd,
PO Box 2008, MS-6017, Oak Ridge, TN 37831, USA
e-mail: gangulyar@ornl.gov

K. Steinhaeuser · N. V. Chawla
Department of Computer Science and Engineering
and Interdisciplinary Center for Network Science
and Applications, University of Notre Dame,
384 Fitzpatrick Hall, Notre Dame, IN 46556, USA

A. R. Ganguly
Department of Civil and Environmental Engineering,
University of Tennessee at Knoxville, 223 Perkins Hall,
Knoxville, TN 37996, USA

2008). While the potential of these tools has been demonstrated and exploited across complex systems in nature and society (Barabási and Bonabeau 2003; Watts and Strogatz 1998) the applications to climate are still emerging (Donges et al. 2009a, b). Availability of massive datasets, whether observations, reanalysis or climate model simulations, has led to new challenges and opportunities. The analysis and insights presented here illustrate the value of *climate networks*, that is, complex networks constructed from climate data.

We describe an adaptation of climate networks constructed from reanalysis-based atmospheric variables (Kalnay et al. 1996) over the oceans, with a focus on identifying dependence structures and their temporal evolutions over varying resolutions in space, within and among multiple variables. The overall network topology (Donges et al. 2009b; Tsonis and Roebber 2004; Tsonis et al. 2006) expresses global properties, while employing community detection to partition the networks (Mucha et al. 2010; Newman 2003; Pons and Latapy 2006; Steinhaeuser and Chawla 2010) (Electronic Supplementary Material [ESM]) reveals additional structure at regional to local scales. Our findings re-confirm known physics-based associations, thus implicitly affirming the validity of our approach, but also suggest new insights in ocean meteorology including the possibility of long-range dependence in atmospheric convective activity. Examining the network dynamics suggests that patterns and dependencies are relatively stable over time, further increasing confidence in our observations.

We develop multivariate and multiscale dependence structures in climate networks from seven variables (Materials and Methods), whereas prior work has mostly focused on univariate analysis or comparisons among a few variables. The spatial proximity between vertices is not explicitly used during network construction. However, both proximity-based correlations and long-range spatial dependence, if any, are expected to emerge from the network structure. The frequency of edge lengths is plotted as a function of spatial distance between the vertices. Climate networks exhibiting small-world properties would suggest a balance between both proximity-based and long-range dependence. A community detection algorithm (Pons and Latapy 2006) is used to partition each network into clusters or regions (see ESM). The important difference from standard clustering algorithms is the use of network distances based on the constructed climate networks, rather than the use of Euclidean or other geography-based distances. The formation of spatial clusters suggests the emergence of patterns based on spatial proximity, whereas clusters that are geographically separated suggest teleconnections. Network and cluster properties based on all available data capture the time-averaged dependence patterns. The temporal evolution of global and regional dependence structures is examined by comparing network and cluster properties over multiple time windows. Evolution in the properties of global climate networks is measured through changes in the edge frequency distribution over time as well as in terms of the ability to predict this distribution for unseen data; evolution of the clusters is quantified through the Adjusted Rand Index (ARI) (Steinhaeuser and Chawla 2010), which is also used to quantify the degree of closeness of structures among multiple variables.

2 Materials and methods

The climate networks developed here rely on the NCEP/NCAR Reanalysis Project (Kalnay et al. 1996), which reconstructed 60 years (1948–2007) of climate data by assimilating remote and in situ sensor measurements across the globe with a physically-based meteorological model. In addition to ensuring internal consistency among variables, reanalysis projects reconstruct variables that are not directly observed. Interpretation of the results requires an understanding of the reconstructions. For the dataset considered here (Kalnay et al. 1996), temperature observations are directly assimilated and hence remain close to observations while precipitable water is derived and hence resemble model outputs more than observations. Measurements are rare or non-existent for updraft or vertical velocities, even though they are important indicators of convective activity (Zelinka and Hartmann 2009). The credibility of reconstructed data relate to both the quality of observations and the physics embedded within models. Thus, insights gained from reanalysis datasets may be compared with observations and tested on climate model simulations. The latter is important for the possible use of the insights found in this study in the context of longer-term climate projections.

Our investigation focuses on the nature of the information content and dependence structures among oceanic variables at the water surface or in the atmosphere. Thus, based on the stated quality and relevance of the variables available from the NCEP/NCAR reanalysis (Kalnay et al. 1996), we selected the following seven for our analysis: sea surface temperature (SST; water temperature at the surface), sea level pressure (SLP; air pressure at sea level), geopotential height (GH; elevation of the 500 mbar pressure level above the surface), precipitable water (PW; vertically integrated water column over the entire atmospheric column), relative humidity (RH; saturation of humidity above the surface), horizontal wind speed (HWS; measured in the plane near the surface), and vertical wind speed (VWS; measured in the atmospheric column).

Since seasonality tends to dominate the climate signal, we consider monthly *anomaly series* for each variable: at each grid point, we calculate for every month (i.e., separately for all Januaries, Februaries, etc.) the long-term mean and standard deviation. Each data point is then normalized by subtracting the mean and dividing by the standard deviation of the corresponding month. This normalization significantly reduces temporal autocorrelation in the time series (Steinbach et al. 2003).

To construct the climate networks, each grid cell is represented by a vertex and weighted edges are created between all pairs of vertices based on the statistical relationship between them (Tsonis and Roebber 2004). The similarity measure used is the cross-correlation between the monthly anomaly series. Because inverse relationships are equally relevant in the present application, we set the edge weight to the absolute value of the correlation coefficient. While nonlinear relationships known to exist in climate might suggest the use of a nonlinear correlation measure, other researchers examined this question and concluded that, "the observed similarity of Pearson correlation and mutual information networks can be considered statistically significant" (Donges et al. 2009a). Thus it seems reasonable to use the simplest possible measure, namely linear (Pearson) correlation. Finally, significance-based pruning is applied to the networks. Specifically, two vertices are considered connected only if the *p-value* of the corresponding correlation is less than 1e-10. This may seem like a stringent requirement but a large number of edges satisfy this criterion and are therefore retained in the networks.

The edge lengths were computed as the great-circle distance between vertices (centers of the corresponding grid cells). For each variable, a histogram using 40 equal-width bins of the edges was computed; for clarity, only lines connecting the mid-points of each bin are shown in the profile plots (Figs. 1a, 2). The maximum length of 20,000 km derives from the fact that this is approximately equal to half of the earth's circumference.

The shape of the profiles (Fig. 1b) is characterized by two properties: the proximity-based spatial autocorrelation is captured by the "peak height", defined as the maximum bin count of the profile; and the long-range spatial dependence is quantified by the "tail thickness", calculated as the area under the profile plot for distances greater than 10,000 km. We selected this threshold based on visual inspection of the profiles to describe the tail but avoid capturing any of the proximity-based autocorrelation. While it is possible to conceive of more principled approaches to threshold selection, empirical evidence suggests that our results and interpretation are not sensitive to the exact threshold chosen.

3 Results

Figure 1a shows the time-averaged dependence patterns at global and regional scales in form of the frequency distribution of edge lengths in the network (called *distance profile*) constructed from each variable. Large values at short distances (a "peak" or a "plateau") indicate the relative dominance of proximity-based spatial correlations. However, non-decaying values at longer distances (a "fat tail") suggest possible dominance of long-range spatial dependence or teleconnections. The dip in frequencies at very short distances (near the *y*-axis) is expected: this is an artifact of the data being arranged on a spherical grid, resulting in increasing spacing when moving from the poles towards the equator and thus having minimum

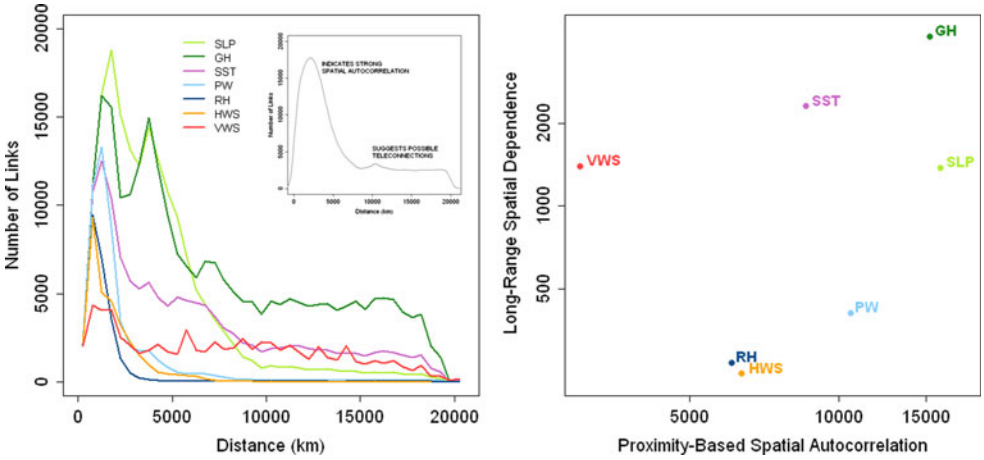

Fig. 1 *Left Panel*: Frequency distributions of the edge lengths in climate networks (distance profiles). Presence of short-distance edges can be interpreted as proximity-based spatial autocorrelation, while long-distance edges may be evidence of long-range spatial dependencies or teleconnections; see inset for an idealized interpretation. *Right Panel*. Visual comparison of the spatial correlation structure in different climate variables based on similarity of the corresponding profiles. The *x*-axis shows a characteristic of the "peak height" and the *y*-axis a characteristic of the "tail thickness" (Materials and Methods), both plotted on a logarithmic scale

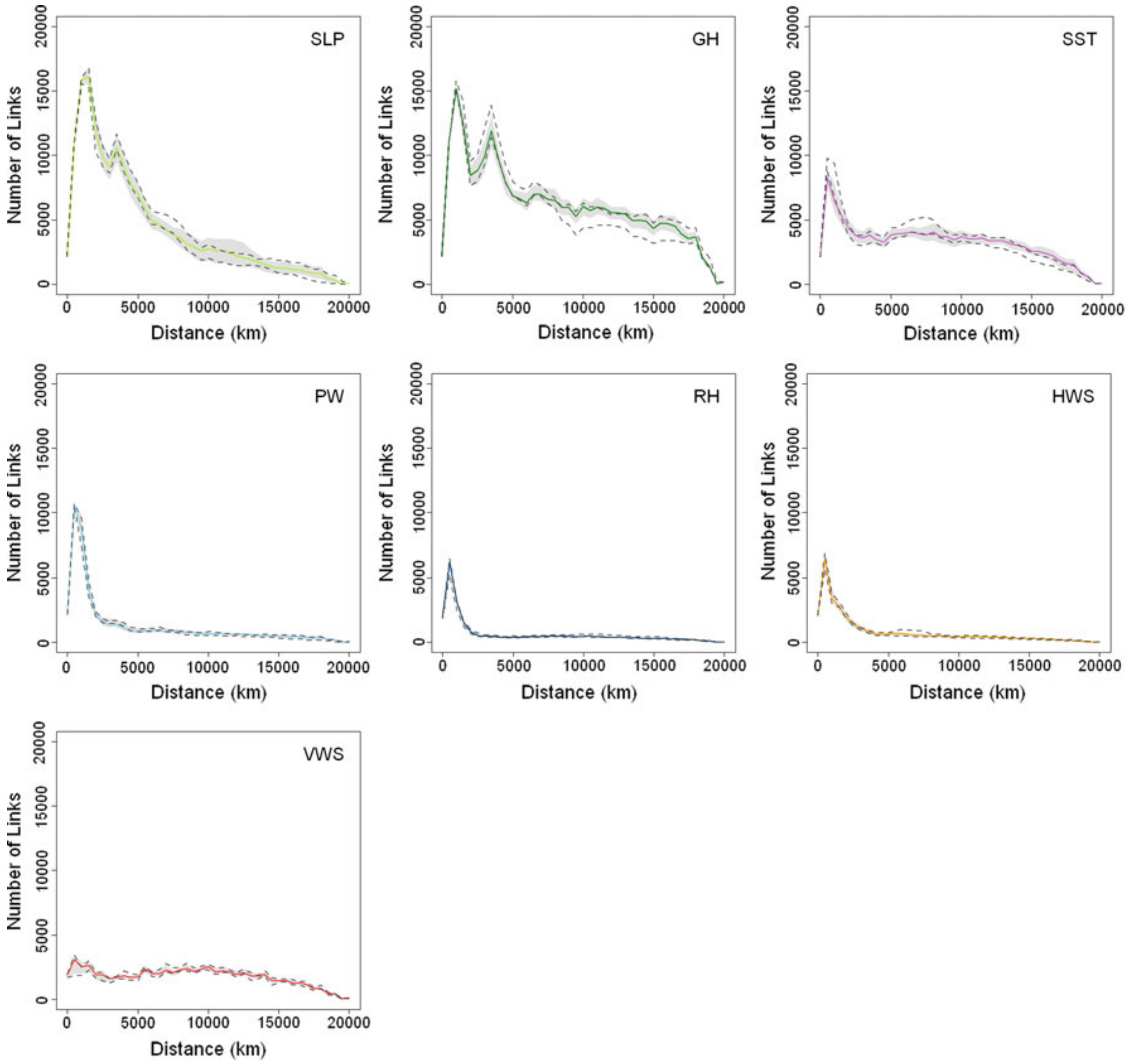

Fig. 2 5th and 95th percentile bounds of the distance profiles, computed separately for each bin over a 40-year baseline period (1952–1991, *shaded*) and the same bounds for a 16-year validation period (1992–2007, *dashed lines*). The fact that the *colored lines* generally fall within the shaded boundaries suggests that the profiles are relatively stable over time

distances greater than the first several histogram bins. Three categories emerge based on the observed profiles: both a high peak and a fat tail—sea surface temperature (SST), sea level pressure (SLP), geopotential height (GH); a moderate peak but no tail—precipitable water (PW), relative humidity (RH), horizontal wind speed (HWS); and near-uniform distribution with a fat tail but little or no peak across the full range—vertical wind speed (VWS). Proximity-based and long-range spatial dependence in SST, SLP and GH follow from basic meteorology (Ahrens 2008) and are well known in climate science; indeed, SST and/or SLP are frequently used to define indices of oceanic oscillators (Alexander et al. 2002; Diaz et al. 2001; Steinbach et al. 2003; Tsonis et al. 2008) and GH is closely correlated with both SST and SLP (Trenberth and Hurrell 1994). Spatial correlation scales of the hydrologic variables are expected to be shorter and there are no known teleconnections, which agrees with the profiles of PW and RH, as well as the similarity between them. Horizontal wind speeds (HWS) are expected to be correlated at relatively short spatial scales but no known teleconnections exists. Vertical wind speeds (VWS) or updraft velocities are an indicator of convective activity, which in turn are known to have very short spatial (and temporal) decorrelation scales (Emanuel 1992). The proximity-based spatial correlation is accordingly rather negligible. However, a fat tail is observed, suggesting the possibility of teleconnections in atmospheric convection patterns over the oceans. This is a surprising observation, which may be a novel insight in climate science or perhaps indicative of a spurious correlation structure from the data. Figure 1b summarizes this categorization with two characteristics, which are assumed to capture the spatial dependence patterns: proximity-based dependence quantified by the "peak height" and teleconnections or long-range spatial dependence quantified by the "tail thickness" (Materials and Methods). The outlying nature of VWS compared to the other variables, which is a consequence of the very short decorrelation scales, becomes immediately apparent. The closeness of RH and HWS dependence behavior is obvious from both panels. While a relation of wind with SST has been observed (O'Neill et al. 2003), this closeness may be of interest for future research.

Figure 2 shows temporal evolution of the dependence structures. First, we construct separate networks for five-year moving windows and calculate their histograms, for a total of 56 profiles per variable. As a baseline, profiles are calculated using the first 40 years of data (1952–1991) and their 5th and 95th percentile bounds are computed. The fact that these bounds are fairly tight suggests that, despite some variability, the dependence structures are fairly stable over time. Second, we use the remaining 16 years (1992–2007) for validation to determine if the bounds of the profiles calculated from the first 40 years have any predictive power for the next 16 years. Indeed, we observe that the validation profiles by and large fall within the bounds, thus lending further credence to temporal stability. A linear regression (von Storch and Zwiers 2002) on each frequency (i.e., histogram bin) shows how the dependence structure develops over time (see ESM: Figures S1 & S2).

Figure 3 shows the multivariate dependence among the clusters produced by partitioning each of the individual networks (see ESM). Cluster similarities are quantified through the Adjusted Rand Index (ARI, see ESM) and the similarity matrix is visualized with a color scheme. The categorization based on the regional or cluster-based dependence is similar to the one based on distance profiles. One category is formed by SST, SLP and GH; a second by PW, RH, and HWS; VWS forms a completely separate category. VWS appears unique compared to the others because of the very short decorrelation scales in space, which is captured through the ARI when compared to the other clusters. As previously noted, the relationship between SST, SLP and GH is well established (Ahrens 2008; Alexander et al. 2002; Diaz et al. 2001; Steinbach et al. 2003; Tsonis et al. 2008). The regional clusters derived from SST and PW show a degree of overlap which is a likely consequence of the Clausius-Clapeyron relation that relates temperature to the water vapor in an atmospheric column (Emanuel 1992; O'Gorman and Schneider 2009). The RH is related to both SST and PW. The relation of HWS to SST (O'Neill et al. 2003) has been suggested in the literature. The close relation of RH and HWS, which is seen from the clusters as well as from the dependence profiles, bears further investigation.

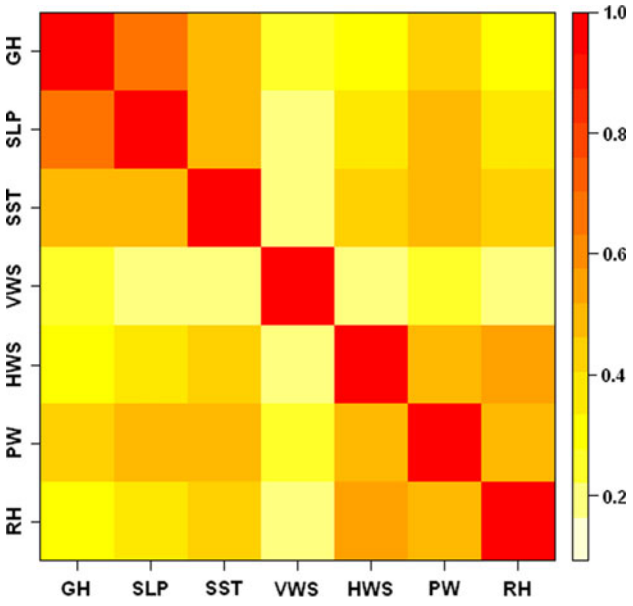

Fig. 3 Visual comparison of the clusterings for different climate variables, calculated as the Adjusted Rand Index (ARI) between their cluster assignments

Figure 4 depicts the oceanic clusters as well as their evolution over time for two 20-year windows, one at the beginning (1948–1967) and another at the end (1988–2007) of reanalysis data availability. We show clusters for several variables but focus our discussion on two in particular. As may be expected, the SST clusters show both proximity-based spatial correlations and teleconnections, many of which are known to meteorologists (Ahrens 2008; Alexander et al. 2002; Diaz et al. 2001; Kumar et al. 2006; Yamasaki et al. 2008), as well as persistence over time. However, the VWS clusters are relatively short-lived with very short correlation scales in space. In this context, the apparent teleconnections in VWS and their persistence over time are rather surprising. The possibility of long-range spatial dependence in atmospheric convective activity and its relationship to other physical processes (e.g., tropical cyclone activity) should be explored in more detail. The ARI is also computed over multiple moving time-windows to quantify the stability of the clusters for each individual variable (see ESM: Figure S3). Stability over time follows a somewhat intuitive pattern: variables that are generally thought of as participating in larger-scale, long term processes are the most stable, whereas variables that participate in highly localized phenomena are more volatile. Somewhat surprisingly, however, PW exhibits higher stability over time than SST. This may be explained by the relatively higher spatial autocorrelation in PW. In addition, we observe statistically significant

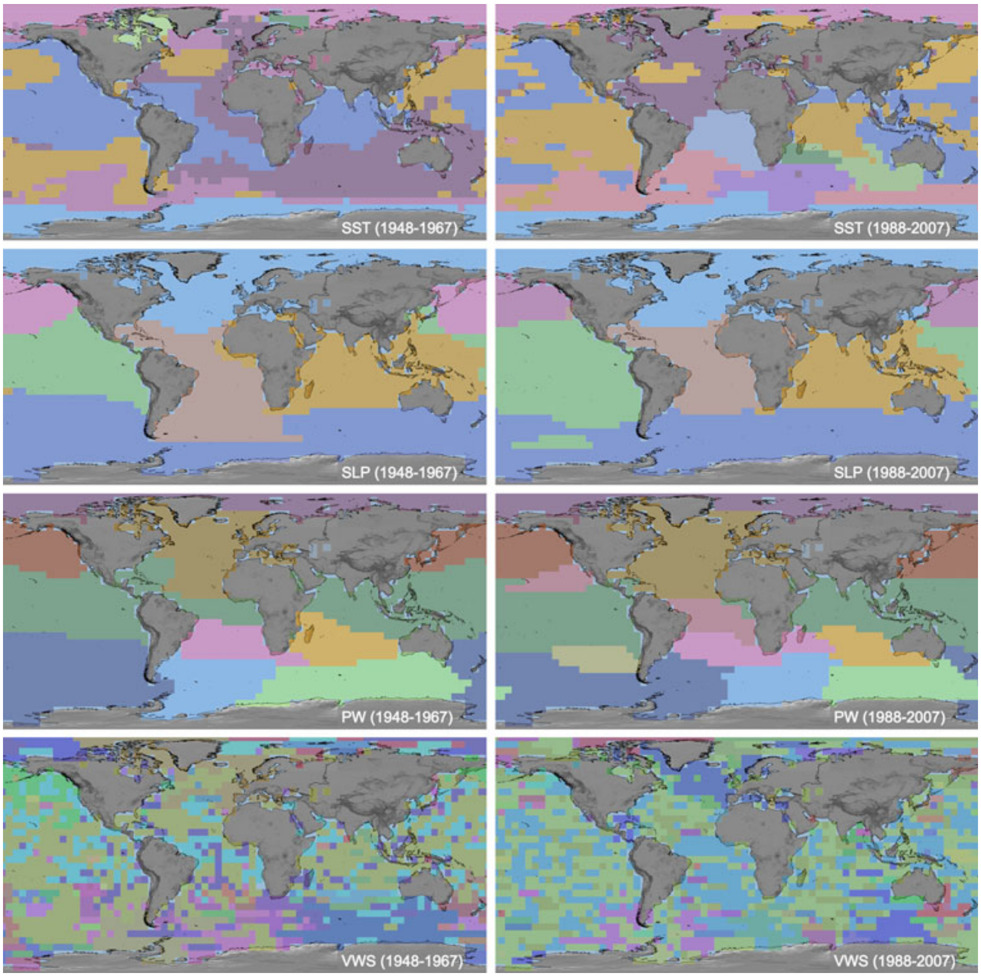

Fig. 4 Network clusters obtained by community detection. Spatial proximity was not used in constructing the networks but some patterns clearly emerge, e.g., spatially coherent clusters for SLP contrasted with relatively scattered clusters for VWS. Networks were constructed from 20-year windows at the beginning (1948–1967) and end (1988–2007) of data availability to illustrate relative stability of these patterns over time

downward trends for both PW and SST, suggesting a destabilization of the network over time; a corresponding upward trend for SLP, suggesting a stabilization of the network; and no significant trends for any of the other variables.

4 Discussion

A recent news article (Schiermeier 2010) claimed that the sad truth of climate science is that model predictions are less reliable at scales and for variables which are most crucial. Advances in climate science or modeling may not keep pace with the urgency of stakeholder requirements in regional climate or hydrologic predictions. However, relatively well-predicted variables like temperature or humidity may have information content for the not so-well predicted but potentially more crucial variables like precipitation. Thus, SST and SLP patterns determine oceanic oscillators which impact regional precipitation over oceans and land (Ahrens 2008; Steinbach et al. 2003), while temperature and humidity profiles in an atmospheric column impacts precipitation extremes over land (O'Gorman and Schneider 2009; Sugiyama et al. 2010). In addition, ancillary variables like updraft velocity may be able to resolve the differences among multiple climate models (O'Gorman and Schneider 2009; Sugiyama et al. 2010). Complex networks may be able to extract the information content in multiple climate variables relevant to a variable of interest, thus leading to better understanding of climate science and offering the possibility of complementing physics-based climate models for improved predictions of the more crucial variables.

Acknowledgments This research was performed as part of a project titled "Uncertainty Assessment and Reduction for Climate Extremes and Climate Change Impacts", which in turn was funded by the initiative called "Understanding Climate Change Impact: Energy, Carbon, and Water Initiative", within the LDRD Program of the Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DE-AC05-00OR22725. Nitesh Chawla was supported in part by the National Science Foundation under Grants OCI-1029584 and BCS-0826958. The United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Ahrens CD (2008) Meteorology today. Brooks Cole, Belmont, CA

Alexander MA et al (2002) The atmospheric bridge: the influence of ENSO teleconnections on air-sea interaction over the global oceans. J Climate 15:2205–2231

Barabási AL, Bonabeau E (2003) Scale-free networks. Sci Am 288:60–69

Diaz JF, Hoerling MP, Eischeid JK (2001) ENSO variability, teleconnections and climate change. Int J Climatol 21:1845–1862

Donges JF, Zou Y, Marwan N, Kurths J (2009a) Complex networks in climate dynamics. Eur Phs J Special Top 174:157–179

Donges JF, Zou Y, Marwan N, Kurths J (2009b) The backbone of the climate network. Europhys Lett 87:48007

Emanuel KA (1992) Atmospheric convection. Oxford University Press, Oxford

Goddard L et al (2001) Current approaches to seasonal-to-interannual climate predictions. Int J Climatol 21:1111–1152

Gozolchiani A, Yamasaki K, Gazit W, Havlin S (2008) Pattern of climate network blinking links follows El Niño events. Europhys Lett 83:28005

Hoerling MP, Kumar A, Xu T (2010) Robustness of the nonlinear climate response to ENSO's extreme phases. J Clim 14:1277–1293

Kalnay E et al (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–470

Kumar KK, Rajagopalan B, Hoerling M, Bates G, Cane M (2006) Unraveling the mystery of Indian monsoon failure during El Niño. Science 314:115–119

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328:876–878

Newman MEJ (2003) Finding and evaluating community structure in networks. Phys Rev E 69:026113

O'Gorman PA, Schneider T (2009) The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proc Nat Acad Sci USA 106:14773–14777

O'Neill LW, Chelton DB, Esbensen SK (2003) Observations of SST-induced perturbations of the wind stress field over the Southern ocean on seasonal timescales, *J*. Climate 16:2340–2354

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Alg App 10:191–218

Schiermeier Q (2010) The real holes in climate science. Nature 463:284–287

Steinbach M, Tan PN, Kumar V, Klooster S, Potter C (2003) Discovery of climate indices using clustering. In: Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining, pp 446–455

Steinhaeuser K, Chawla NV (2010) Identifying and evaluating community structure in complex networks. Pattern Rec Lett 31:413–421

Steinhaeuser K, Chawla NV, Ganguly AR (2010a) An exploration of climate data using complex networks. ACM SIGKDD Explor 12:25–32

Steinhaeuser K, Chawla NV, Ganguly AR (2010b) Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Stat Anal Data Mining (in press)

Sugiyama M, Shiogama H, Emori S (2010) Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models. Proc Nat Acad Sci USA 107:571–575

Trenberth KE, Hurrell JW (1994) Decadal atmosphere-ocean variations in the Pacific. Clim Dynam 9:303–319

Tsonis AA, Roebber PJ (2004) The architecture of the climate network. Physica A 333:497–504

Tsonis AA, Swanson KL (2008) Topology and predictability of El Niño and La Niña networks. Phys Rev Lett 100:228502

Tsonis AA, Swanson KL, Roebber PJ (2006) What do networks have to do with climate? Bull Am Meteorol Soc 87:585–595

Tsonis AA, Swanson KL, Wang G (2008) On the role of atmospheric teleconnections in climate. J Clim 21:2990–3001

von Storch H, Zwiers FW (2002) Statistical analysis in climate research. Cambridge University Press, Cambridge, UK

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Yamasaki K, Gozolchiani A, Havlin S (2008) Climate networks around the globe are significantly affected by El Niño. Phys Rev Lett 100:157–179

Zelinka MD, Hartmann DL (2009) Response of humidity and clouds to tropical deep convection. J Clim 22(9):2389–2404